

Helsinki 28.3.2000

REC'D 28 APR 2000

PO

PCT

ETUOIKEUSTODISTUS
PRIORITY DOCUMENT

4



Hakija
Applicant

Alma Media Oyj
Helsinki

Patenttihakemus nro
Patent application no

990286

Tekemispäivä
Filing date

12.02.1999

Kansainvälinen luokka
International class

G06F

Keksinnön nimitys
Title of invention

"Mekanismi sähköisen tekstihaun tukemiseksi"

Täten todistetaan, että oheiset asiakirjat ovat tarkkoja jäljennöksiä patentti- ja rekisterihallitukselle alkuaan annetuista selityksestä, patenttivaatimuksista, tiivistelmästä ja piirustuksista.

This is to certify that the annexed documents are true copies of the description, claims, abstract and drawings originally filed with the Finnish Patent Office.


Pirjo Kalla
Tutkimussihteeri

**PRIORITY
DOCUMENT**

SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

Maksu 300,- mk
Fee 300,- FIM

Osoite: Arkadiankatu 6 A Puhelin: 09 6939 500 Telefax: 09 6939 5328
P.O.Box 1160 Telephone: + 358 9 6939 500 Telefax: + 358 9 6939 5328
FIN-00101 Helsinki, FINLAND

Mekanismi sähköisen tekstihaun tukemiseksi

Keksinnön tausta

Keksintö liittyy sähköisen tekstihaun tukemiseen, erityisesti kohdistettaessa hakuja Internet-tyyppisessä verkossa ja CD-ROM -levyillä julkaistaviin dokumentteihin.

Sähköisessä muodossa julkaistavien dokumenttien määrä ja informaation sisältö kasvavat valtavalla nopeudella. Yhä enenevä osa artikkeleista julkaistaan Internetissä tai CD-ROM -levyillä (tai DVD-levyillä).

Käyttäjä etsii tietoa tällaisista dokumenteista antamalla yhden tai muutaman sanan, joita hän pitää erityisen relevantteina. Näitä sanoja kutsutaan hakusanoiksi. Esimerkiksi maantieliikenneonnettomuuksista kiinnostunut käyttäjä voi etsiä hakusanoja "tie", "liikenne", "onnettomuus" jne.

Ohjelmaa, tietokonetta ja palvelua, joka toteuttaa käyttäjän määrittelemän tekstihaun, kutsutaan vastaavasti hakuohjelmaksi, hakukoneeksi ja hakupalveluksi. Jos hakuohjelma ensin vastaanottaisi käyttäjän määrittelemät hakusanat ja vasta sitten lähtisi seulomaan koko sen ulottuvilla olevaa informaatiota, haku muodostuisi yleensä toivottoman hitaaksi. Täyteen kirjoitetun CD-ROM -levyn läpikäyminen kestäisi useita minuutteja. Koko Internetin läpikäyminen veisi useita päiviä tai viikkoja. Tämä myös kuormittaisi Internetiä kohtuuttomasti. Koska ensimmäinen haku ei yleensä tuota riittävän hyvää otosta, haku joudutaan toistamaan useita kertoja.

Tämän ongelman ratkaisemiseksi on kehitetty indeksointiohjelmia ja -palveluja, jotka käyvät etukäteen läpi käytettävissään olevaa informaatiota ja muodostavat siitä indeksoidun tietokannan, johon voidaan kohdistaa hakuja yhdessä tai muutamassa sekunnissa. Esimerkkejä tällä tekniikalla Internetissä toimivista hakupalveluista ovat Lycos, Yahoo ja AltaVista. Esimerkkinä omassa tietokoneessa tai lähiverkossa toimivasta hakuohjelmasta olkoon dtSearch, jota valmistaa samanniminen yhtiö. Kaikista näistä on saatavana tietoa www-osoitteella (World Wide Web) www.nimi.com, missä "nimi" on yllä mainittu palvelun tai yhtiön nimi.

Kuvio 1 esittää Internet-tyyppisessä verkossa julkaistavan dokumentin hakua. Internet-tyyppisellä verkolla tarkoitetaan varsinaisen Internet-verkon lisäksi sen suljettuja osa-alueita, joista käytetään nimityksiä intranet, extranet jne. TE (Terminal Equipment) esittää käyttäjän päätelaitteistoa, jolla tarkoitetaan käyttäjän tietokonetta ja/tai näyttöpäätettä sekä siinä suoritettavaa selainohjelmaa Internet-sivujen esittämiseksi. Viite 1-A esittää hakupalvelun

tarjoajan hakupalvelinta, DNS (Domain Name Server) yhtä nimipalvelun palvelinta eli nimipalvelinta ja viite 1-B Internet-sivuja ylläpitävää WWW-palvelinta eli dokumentin julkaisijan palvelinta.

Vaiheessa 1-2 käyttäjän antama hakupalvelimen HTTP-muotoinen
5 (HyperText Transfer Protocol) Internet-osoite välitetään nimipalvelimelle DNS, joka puolestaan välittää käyttäjälle vaiheessa 1-4 kyseisen hakupalvelimen osoitteen IP-muodossa (Internet Protocol). IP-osoitteen avulla päätelaitteisto TE muodostaa vaiheessa 1-6 yhteyden hakupalvelimeen 1-A.

Vaiheessa 1-8 hakupalvelin lähettää WWW-sivunsa käyttäjälle si-
10 vunkuvauskielen HTML-muodossa (HyperText Markup Language), ja WWW-sivut esitetään käyttäjän päätelaitteiston TE näyttöpäätteellä. Yhteys palvelimen ja päätelaitteiston välillä on auki vain sivun siirtoon kuluvan ajan.

Vaiheessa 1-10 käyttäjä antaa hakukoneen hakulomakkeeseen yh-
den tai useampia hakusanoja, ja ne lähetetään vaiheessa 1-12 hakupalveli-
15 melle. Hakuohjelma etsii vaiheessa 1-14 kyseisiä hakusanoja hakupalvelimen tietokannasta. Lista löytyneistä, hakusanat sisältävistä dokumenteista palautetaan käyttäjälle vaiheessa 1-16.

Vaiheessa 1-18 käyttäjä voi selata löytyneitä dokumentteja Internet-selaimessaan. Kun hän haluaa tutustua johonkin haussa löytyneeseen doku-
20 menttiin, hän antaa kyseisen dokumentin WWW-osoitteen selaimelleen (esimerkiksi valitsemalla sen hakupalvelun tuottamasta listasta), joka ottaa yhteyden vaiheessa 1-20 nimipalvelimeen. Tämä palauttaa vaiheessa 1-22 kyseisen dokumentin IP-osoitteen selaimelle, joka tämän IP-osoitteen avulla pyytää kyseistä dokumenttia vaiheessa 1-24. Vaiheessa 1-26 kyseinen WWW-
25 sivu lähetetään käyttäjälle. Mikäli käyttäjä haluaa selata muita hakukoneen löytämiä dokumentteja, hän voi vaiheessa 1-28 palata takaisin hakukoneen listaukseen löytyneistä dokumenteista.

Käyttäjä voi toistaa vaiheita 1-18 ... 1-28, kunnes hän on käynyt läpi
30 kaikki hakukoneen löytämät dokumentit ja/tai kunnes hän haluaa lopettaa dokumenttien selaamisen.

Jotta edellä kuvattu haku olisi tehokasta, eri hakupalveluilla on erilaisia tekniikoita hakusanojen yhdistelemiseksi. Yleisesti käytetään loogisia operaattoreita AND, OR ja NOT sekä sulkumerkkejä. Esimerkiksi haku sanoilla
35 "tie AND onnettomuus" etsii dokumentteja, joissa esiintyvät sanat "tie" ja "onnettomuus".

Haku ei kuitenkaan yleensä tuota relevanttia tietoa, mikäli annetaan vain hakusanojen yhdistelmä. Sen vuoksi useimmat hakupalvelut tunnistavat myös läheisyysoperaattorin. Altavistan tapauksessa tämä on NEAR. Haku sanoilla "tie NEAR onnettomuus" etsii dokumentteja, joissa sanat "tie" ja "onnettomuus" esiintyvät korkeintaan 10 sanan etäisyydellä toisistaan. Myös dtSearch -ohjelmalla voidaan määritellä maksimaalinen sanojen etäisyys: läheisyysoperaattori w/n, missä $n=1, 2, \dots$, edellyttää että sanat esiintyvät korkeintaan $n:n$ sanan etäisyydellä toisistaan.

Tämän keksinnön perustana oleva ongelma on, että etukäteen tapahtuva indeksointi toimii huonosti kielissä, joissa sanoilla on useita taivutusmuotoja. Suomen kielen nomineilla ja verbien nominaalimuodoilla on 15 sijamuotoa, unkarin kielessä peräti 21. Kun otetaan huomioon yksikkö- ja monikkomuodot, possessiivisuffiksit ja muut päätteet, mahdollisia taivutusmuotoja on useita satoja.

Taivutettujen muotojen löytämiseksi kaikki yllä mainitut hakupalvelut tukevat villi- eli tähtimerkin (*) käyttöä: tähtimerkkiä voidaan käyttää osoittamaan, että sanan loppu on katkaistu ja hakupalvelun tulee löytää kaikki sanat, jotka alkavat annetulla tavalla. Esimerkiksi annettaessa hakusanaksi "onnettomuu" tulee hakupalvelun löytää sanat "onnettomuudet", "onnettomuuksista", "onnettomuustutkintalautakunta" jne.

Tähtimerkin käytössä on kuitenkin ongelmia ja rajoituksia. Esimerkiksi AltaVista -hakupalvelu vaatii, että hakusanasta annetaan ainakin kolme kirjainta ennen tähtimerkkiä. Kuitenkin esimerkiksi sanan "tie" taipumaton osa on vain yhden kirjaimen pituinen: "teiden", "teillä" jne. Toinen ongelma on, että hakusanalla "tie*" palautetaan kaikki tie-alkuiset sanat, kuten "tiede", "tietokone", "tietoliikenne", "tietysti", "tienoo" ja "tietoisuus", kaikissa taivutusmuodoissaan. Lyhytvartaloisten sanojen etsiminen tunnetulla tekniikalla tuottaa siis erittäin epärelevanttia tai ei lainkaan tietoa.

Keksinnön lyhyt selostus

Keksinnön eräänä tarkoituksena on kehittää sellainen sähköisen dokumentin rakenne, jolla dokumentin indeksoinnin jälkeen ei esiinny yllä mainittuja ongelmia. Toisella tavalla nähtynä keksinnön tavoitteena on kehittää menetelmä ja laitteisto tällaisten dokumenttien tuottamiseksi. Keksinnön tavoitteet saavutetaan menetelmällä ja järjestelmällä, joille on tunnusomaista se, mitä sanotaan itsenäisissä patenttivaatimuksissa. Keksinnön edulliset suoritusmuodot ovat epäitsenäisten patenttivaatimusten kohteena.

Keksintö perustuu siihen, että sähköisesti julkaistavaa dokumenttia täydennetään lisäämällä siihen dokumentin tekstiosuuden sisältämät sanat perusmuodoissaan ja alkuperäisessä järjestyksessä. Sanojen lisääminen perusmuodoissaan saa aikaan sen, että hakupalvelu löytää seuraavan indeksoinnin jälkeen keksinnön mukaisesti täydennetyn dokumentin, vaikka alkuperäisessä dokumentissa sana ei esiintyisi lainkaan perusmuotoisena.

Jäljempänä käytetään nimitystä "täydennysosa" siitä osasta, joka sisältää keksinnön mukaisesti lisätyt sanat. Vastaavasti "perusosa" on se osa, joka sisältää alkuperäisen dokumentin.

Itse asiassa on tunnettua lisätä dokumentteihin käsin joitakin perusmuotoisia avainsanoja. Tieteellisten dokumenttien otsikon alla tai vaihtoehtoisesti dokumentin lopussa käytetään joskus kenttää "avainsanat", jossa esiintyy muutama avainsana. Tämä ei kuitenkaan ratkaise ongelmaa toivotulla tavalla, koska perusmuotoisten avainsanojen määrä on hyvin rajallinen, eikä läheisyysoperaattori toimi oikein. Pitkässä artikkelissa voidaan puhua useasta täysin erillisestä asiasta, mutta avainsanakentässä vastaavat hakusanat ovat kuitenkin lähellä toisiaan.

Keksinnön mukainen tekniikka, jossa sanat lisätään alkuperäisessä järjestyksessä saa aikaan sen, että hakupalvelu osaa käyttää oikein läheisyysoperaattoreita. Esimerkiksi haku sanoilla "tie NEAR onnettomuus" löytäisi dokumentin, joissa esiintyy tekstifragmentti "teilläämme tapahtuneet onnettomuudet", vaikka dokumentti ei sisältäisi lainkaan sanoja "tie" tai "onnettomuus" perusmuodoissaan.

Koska dokumentin tekstiosan sanat lisätään alkuperäisessä järjestyksessä, näyttäisi siltä että dokumentin pituus likimain kaksinkertaistuu. Tämä pitää paikkansa vain tekstiä sisältävien dokumenttien suhteen. Useimpiin dokumentteihin liittyy kuitenkin kuvia, joiden vaatima muistitila ylittää moninkertaisesti tekstiosuuden vaatiman muistitilan, joten tekstiosuuden kaksinkertaistaminen ei merkittävästi kasvata koko dokumentin vaatimaa muistitilaa.

Muistitilan vähäisen kasvamisen vastapainoksi keksinnön mukainen tekniikka tuo vielä yhden yllättävän edun: näin täydennettyjen dokumenttien relevanssi kasvaa näennäisesti ainakin kaksinkertaiseksi, koska dokumenteissa on käyttäjän valitsemia hakusanoja kaksinkertainen määrä. Keksinnön mukaisesti täydennetyn dokumentin julkaisija saa siis sanomansa paremmin perille. Dokumentin relevanssi kasvaa kaksinkertaiseksi sellaisten hakusanojen suhteen, joilla on niin pitkä vartalo, että niitä voidaan luotettavasti hakea tähti-

merkillä, esimerkiksi "onnettomuu*". Lyhytvartaloisten sanojen kohdalla, joita tunnetulla tekniikalla ei voida hakea lainkaan, dokumentin relevanssi kasvaa moninkertaiseksi, mikä johtuu siitä, että tunnetulla tekniikalla tällaiset dokumentit eivät hakupalvelulle ole lainkaan relevantteja. (Ne voivat olla osittain relevantteja siinä tapauksessa, että käyttäjä antaa useita hakusanoja, joista muut sanat ovat sellaisia, että hakupalvelu löytää ne.) Tässä kappaleessa relevanssilla ei siis tarkoiteta sitä, kuinka relevantti jokin dokumentti on käyttäjälle, mikäli hän sen löytäisi, vaan sillä tarkoitetaan hakupalvelun tuottamaa mittalukua, jonka laskenta perustuu siihen, kuinka monta annetuista hakusanoista esiintyy dokumentissa, ja mahdollisesti kuinka usein ne esiintyvät.

Dokumentin käyttäjät (henkilöt, jotka etsivät kyseistä dokumenttia) eivät voi etukäteen tietää, mitkä dokumentit on täydennetty keksinnön mukaisella tavalla, ja mitkä eivät ole. Myös tästä syystä sanojen lisääminen alkupe-
 15 ~~raitse muuttaa hakutottumuksiaan mitenkään, vaan he voivat käyttää läheis-~~
 syysoperaattoria totutulla tavalla.

Käyttäjän hakutoiminto ei kuitenkaan lopu siihen, että hakupalvelu löytää hänelle jonkin hakusanat sisältävän dokumentin. Hänen on yleensä vielä löydettävä relevantit alueet dokumentin sisältä.

20 Oletetaan aluksi, että dokumentit täydennetään yksinkertaisesti lisäämällä perusmuotoiset sanat dokumentin loppuun. Käyttäjä voi etsiä tästä täydennysosuudesta perusmuotoisia hakusanoja selain- tai tekstinkäsittelyohjelman hakutoiminnoilla. Mikäli hakusana on lyhytvartaloinen, käyttäjä ei voi etsiä sitä dokumentin perusosasta, mutta hän voi katsoa täydennysosasta
 25 jonkin hakusanan lähellä olevan pidemmän ja harvinaisemman sanan, ja etsiä sen dokumentin perusosasta. Tässä suhteessa keksinnön mukainen tekniikka voi aiheuttaa pienen muutoksen käyttäjän toimintatapoihin, mutta muutos näkyy vasta sitten kun hakuohjelma on jo löytänyt dokumentin ja käyttäjä selaa sitä. Mikäli dokumentti on lyhyt, tai käyttäjä muusta syystä päättää lukea sen
 30 kokonaan, käyttäjä ei joudu muuttamaan toimintatapojaan.

Dokumentin perusmuotoisten sanojen lisääminen dokumentin loppuun vääristää dokumentin ulkoasua. Teksti näyttää sellaisen ihmisen kirjoittamalta, joka ei ymmärrä kielestä mitään, vaan kääntää koneellisesti sanakirjan avulla. Dokumentin kirjoittajan mielestä tällaista voitaisiin pitää jopa res-
 35 pektioikeuden loukkauksena. (Respektioikeus tarkoittaa, että kaupallisesta levitysoikeudesta riippumatta teosta ei saa esittää loukkaavalla tavalla.) Sen

vuoksi täydennysosa on edullista liittää dokumenttiin tavalla, joka estää sen näkymisen dokumentin normaalikäytössä. Esimerkiksi HTML-koodattuun (HyperText Markup Language) dokumenttiin voidaan liittää ainakin yksi kommentti- tai metakoodikenttä, joka sisältää keksinnön mukaisen täydennysosan.

- 5 Vaihtoehtoinen tapa on yhden tai useamman kuvan lataaminen täydennysosan päälle. Kun käyttäjä haluaa etsiä hakusanaa tästä täydennysosasta, hän avaa dokumentin selainohjelmallaan ja näyttää dokumentin sisältämät HTML-kieliset käskyt. Esimerkiksi Internet Explorer -ohjelmalla tämä tapahtuu käskyllä View/Source. Vastaavasti kehittyneillä tekstinkäsittelyohjelmilla on mahdollista asettaa täydennysosalle attribuutti "piiloteksti", jolloin se saadaan nä-
10 kyviin näyttämällä normaalisti näkymättömät ohjaus- ja erikoismerkit.

- Tunnetut hakutekniikat eivät löydä hakusanoja, jotka esiintyvät yhdyssanan osina, mutta eivät sen alussa. Läheisyysoperaattori ei myöskään toimi, mikäli hakusanat esiintyvät yhdyssanan eri osina. Esimerkiksi sanasta
15 "maantieliikenteen" ei löydetä sanoja "tie" eikä "liikenne" eikä varsinkaan näitä sanoja lähellä toisiaan. Sen vuoksi-erään-toisen-edullisen suoritusmuodon mukaan keksinnön mukainen täydennysosa sisältää kunkin yhdyssanan kohdalla kyseisen yhdyssanan perusmuodon lisäksi yhdyssanan osien perusmuodot erillisinä sanoina. Esimerkiksi taivutetussa muodossa olevan yhdyssanan
20 "maantieliikenneonnettomuuksien" kohdalla täydennysosa sisältäisi sanat "maantieliikenneonnettomuus", "maa", "tie", "liikenne" ja "onnettomuus". Näin täydennetty dokumentti löytyy, mikäli käyttäjä hakee sanoja "tie" ja "liikenne", jopa silloin kun käyttäjä vaatii, että nämä sanat esiintyvät lähekkäin.

- Vielä erään edullisen suoritusmuodon mukaan keksinnön mukainen
25 täydennysosa sisältää perusmuotoisen yhdyssanan ja sen osien lisäksi kaikki yhdyssanan osien yhdistelmät siten, että yhdyssanan muut kuin viimeinen osa ovat siinä muodossa kuin ne esiintyvät dokumentissa ja yhdyssanan viimeinen osa on perusmuodossaan. Yhdyssanan osien yhdistelmät ovat lisäksi alkuperäisessä järjestyksessään, siis edellisen esimerkin tapauksessa sanat
30 "maantie", "tieliikenne" ja "liikenneonnettomuus" sekä "maantieliikenne", ja "tieliikenneonnettomuus".

- Vielä erään edullisen suoritusmuodon mukaan keksinnön mukainen täydennysosa lisää keskitetyssä palvelimessa, jotta jokaisen dokumentin julkaisijan ei tarvitsisi hankkia ohjelmistoa, joka osaa muuntaa sanoja perus-
35 muotoonsa. Alkuperäinen dokumentti voidaan lähettää täydennystä varten levykkeellä, sähköpostin liitetiedostona, Internetin FTP-protokollalla tms.

Kuvioiden lyhyt selostus

Keksintöä selostetaan nyt lähemmin edullisten suoritusmuotojen yhteydessä, viitaten oheisiin piirroksiin, joista:

5 Kuvio 1 on yhdistetty vuo- ja signaalointikaavio, joka esittää Internet-tyyppisessä verkossa julkaistavan dokumentin hakua;

Kuvio 2 esittää signaalointikaaviota keksinnön mukaisen täydennysosan lisäämiseksi;

Kuvio 3A esittää esikäsitellyn dokumentin ja täydennetyn dokumentin rakenteita.

10 Kuvio 3B esittää täydennetyn dokumentin rakennetta, missä täydennysosan päälle on ladattu kuva.

Keksinnön yksityiskohtainen selostus

15 Eräs mahdollinen tekniikka keksinnön mukaisen täydennysosan lisäämiseksi dokumenttiin esitetään kuviossa 2, jossa viite 2-A esittää dokumentin julkaisijan palvelinta, DNS nimipalvelinta, viite 2-B edellä mainittua keskitettyä palvelinta eli täydennyspalvelun tuottajan palvelinta ja viite 2-C hakupalvelun tarjoajan palvelinta

20 Dokumentin julkaisijalla tarkoitetaan sitä, joka haluaa julkaista keksinnön mukaisesti täydennetyn dokumentin. Täydennyspalvelun tuottaja puolestaan tarjoaa keksinnön mukaisen palvelun täydennysosan lisäämiseksi dokumenttiin.

25 Vaiheessa 2-10 dokumentin julkaisija lähettää täydennyspalvelun tuottajan palvelimen WWW-osoitteen nimipalvelimelle DNS, joka palauttaa vaiheessa 2-12 vastaavan IP-osoitteen dokumentin julkaisijalle. Tämän avulla dokumentin julkaisija pääsee vaiheessa 2-14 täydennyspalvelun tuottajan Internet-sivuille. Vaiheessa 2-16 dokumentin julkaisijan selainohjelma noutaa täydennyspalvelun tuottajan WWW-sivun/-sivut päätelaitteelleen.

30 Kyseisellä WWW-sivulla voidaan esittää ainakin sähköpostiosoite, jonne dokumentin julkaisija voi lähettää dokumentin keksinnön mukaista täydennystä varten. Sivulla täydennyspalvelun tuottaja voi kertoa esimerkiksi tarjoamastaan palvelusta ja antaa ohjeita sen käyttämiseksi.

35 Saatuaan sähköpostiosoitteen tietoonsa dokumentin julkaisija voi vaiheessa 2-18 lähettää dokumenttinsa muokattavaksi täydennyspalvelun tuottajalle liittämällä sen esimerkiksi sähköpostin liitetiedostoksi (attachment).

Toinen mahdollinen dokumentin lähetystapa on FTP-siirto (File Transfer Protocol). Dokumentin siirtotapa ei kuitenkaan ole keksinnön kannalta oleellinen.

Vaiheessa 2-20 dokumentin julkaisijan HTML-muotoista dokumenttia muokataan täydennyspalvelun tuottajan palvelimella: siihen lisätään keksinnön mukainen täydennysosa. Tämän jälkeen kyseinen täydennetty dokumentti palautetaan dokumentin julkaisijalle vaiheessa 2-22 joko sähköpostilla tai FTP:n avulla. Sen jälkeen kun hakupalvelu on indeksoinut vaiheessa 2-24 kyseisen dokumentin, dokumentin käyttäjä voi etsiä kyseistä dokumenttia myös perusmuotoisilla sanoilla ja, mikäli yhdyssanat jaetaan osiin, myös niiden osien perusmuodoilla. Internet-tyyppisessä verkossa julkaistavan dokumentin hakua kuvataan kuvion 1 vaiheesta 1-12 alkaen.

Hakupalvelin on konfiguroitavissa siten, että täydennysosan perusmuotoisille sanoille voidaan antaa myös enemmän painoarvoa eli relevanssipisteitä kuin normaalisti. Tämä tarkoittaa sitä, että dokumentit, joiden sisältämillä sanoilla on enemmän relevanssipisteitä sijoitetaan hakutulokset esittävässä listassa lähemmäksi listan kärkipäätä kuin dokumentit, joiden sisältämillä sanoilla on vähemmän relevanssipisteitä. Jos täydennysosan perusmuotoisille sanoille ei anneta ollenkaan relevanssipisteitä tai jos sanat jätetään indeksoimatta, dokumenttia ei löydetä täydennysosan avulla.

Kuvio 3A esittää esikäsitellyn dokumentin 3-2 ja täydennetyn dokumentin 3-20 rakenteita. Alkuperäinen dokumentti voi olla esimerkiksi tekstin käsittelyohjelmalla kirjoitettu tekstisivu muotoiluineen. Esikäsitelty dokumentti on alkuperäinen dokumentti muokattuna esimerkiksi HTML-kieliseksi. Dokumentteissa voi olla lisäksi kuvia, taulukoita, kehyksiä ja/tai muita Internet-sivuilla tallennettavissa olevia objekteja. Viite 3-10 kuvaa HTML-kielen aloitusmerkkiä <HTML> ja viite 3-12 kuvaa HTML-kielen lopetusmerkkiä </HTML>. Kyseisten merkkien välissä on dokumentin sisältö 3-4.

Keksinnön mukaisesti täydennetty dokumentti 3-20 sisältää myös aloitusmerkin 3-10 ja lopetusmerkin 3-12 sekä dokumentin sisällön 3-4. Tämän lisäksi täydennettyyn dokumenttiin 3-20 on lisätty täydennysosa 3-24, jossa kaikki dokumentissa esiintyvät sanat ovat perusmuodoissaan alkuperäisessä järjestyksessä. Täydennysosan koodaus voi tapahtua esimerkiksi koodaamalla se metakoodiksi (Metakeyword) tai HTML-komentiksi. HTML-tiedosto voi sisältää useita HTML-kommentteja. HTML-tiedoston erotinmerkkeinä toimivat "<! ..." ja "...>". Kommentin sijainnilla tiedostossa ei ole merkitystä. Kommentti voi olla dokumentin (3-20) alussa, lopussa tai sen keskellä. Näiden

tekniikoiden sijasta tai niiden lisäksi täydennysosan 3-24 päälle voidaan ladata yksi tai useampi kuva. Tätä täydennetytyn dokumentin 3-40 rakennetta esitetään kuviossa 3B. Kun täydennysosa 3-24 on kuvan 3-44 alla, täydennysosa ei tule näkyviin dokumentin normaalikäytössä. Tällöin näkyvillä on ai-
5 noastaan esikäsitelty dokumentti 3-42.

Perusmuotoisten sanojen lisäksi dokumentin täydennysosaan voidaan lisätä myös sanojen eri variaatioita, synonyymejä ja rinnakkaismerkityksiä. Tällöin dokumentin relevanssi kasvaa edelleen, koska dokumenttia voidaan etsiä myös hakusanoilla, joita ei esiinny alkuperäisessä dokumentissa.

10 Alan ammattilaiselle on ilmeistä, että tekniikan kehittyessä keksinnön perusajatus voidaan toteuttaa monin eri tavoin. Keksintö ja sen suoritusmuodot eivät siten rajoitu yllä kuvattuihin esimerkkeihin vaan ne voivat vaihdella patenttivaatimusten puitteissa.

Patenttivaatimukset

1. Menetelmä ainakin tekstiosuuden sisältävän dokumentin (3-20, 3-40) julkaisemiseksi yhdelle tai useammalle käyttäjälle, jossa menetelmässä:

- dokumentin (3-20, 3-40) tosiaikaisen haun tehostamiseksi dokumenttiin (3-20, 3-40) kohdistetaan ainakin yksi indeksointi (2-24) ja
- indeksoinnin tulos tallennetaan,

tunnettu siitä, että ennen mainittua ainakin yhtä indeksointia (2-24) dokumenttia (3-20, 3-40) täydennetään lisäämällä (2-20) siihen täydennysosa (3-24), joka sisältää olennaisesti ainakin kyseisen dokumentin (3-20, 3-40) tekstiosuuden sisältämät sanat perusmuodoissaan alkuperäisessä järjestyksessä.

2. Patenttivaatimuksen 1 mukainen menetelmä, tunnettu siitä, että täydennysosa (3-24) liitetään dokumenttiin (3-20, 3-40) tavalla, joka estää sen näkymisen dokumentin normaalikäytössä.

3. Patenttivaatimuksen 1 mukainen menetelmä, tunnettu siitä, että täydennysosa (3-24) sisältää kunkin yhdyssanan kohdalla kyseisen yhdyssanan perusmuodon lisäksi yhdyssanan osien perusmuodot erillisinä sanoina.

4. Patenttivaatimuksen 3 mukainen menetelmä, tunnettu siitä, että täydennysosa (3-24) sisältää lisäksi kaikki yhdyssanan osien yhdistelmät, missä osat ovat alkuperäisessä järjestyksessä.

5. Patenttivaatimuksen 1 mukainen menetelmä, tunnettu siitä, että useita dokumentteja julkaistaan usealla julkaisupalvelimella ja että täydennysosa (3-24) lisätään täydennyspalvelimella (2-B), joka on yhteinen usealle julkaisupalvelimelle.

6. Patenttivaatimuksen 5 mukainen menetelmä, tunnettu siitä, että täydennyspalvelin (2-B) vastaanottaa (2-18) ja lähettää (2-22) täydennettävät dokumentit (3-20, 3-40) IP-protokollaa käyttävän tietoliikenneverkon kautta.

7. Laitteisto (2-B) elektronisen tekstihaun tukemiseksi, joka laitteisto (2-B) on sovitettu vastaanottamaan ainakin tekstiosuuden sisältämän dokumentin (3-20, 3-40), tunnettu siitä, että dokumentin (3-20, 3-40) tosiaikaisen haun tehostamiseksi laitteisto (2-B) on sovitettu lisäämään täydennysosan (3-24), joka sisältää olennaisesti ainakin kyseisen dokumentin (3-20, 3-40) tekstiosuuden sisältämät sanat perusmuodoissaan alkuperäisessä järjestyksessä.

8. Patenttivaatimuksen 7 mukainen laitteisto (2-B), tunnettu siitä, että laitteisto (2-B) on sovitettu vastaanottamaan (2-18) ja lähettämään (2-22) dokumentti (3-20, 3-40) IP-protokollaa käyttävän tietoliikenneverkon kautta.

5 9. Järjestely dokumenttien julkaisemiseksi IP-protokollaa käyttävän tietoliikenneverkon kautta, joka järjestely käsittää ainakin yhden julkaisupalvelimen (1-B, 2-A) mainitun dokumentin julkaisemiseksi, ainakin yhden indeksointipalvelimen (1-A, 2-C) mainitun dokumentin indeksoimiseksi ja ainakin yhden päätelaitteen (TE) kyselyn lähettämiseksi mainitulle ainakin yhdelle in-
10 deksointipalvelimelle (1-A, 2-C), tunnettu siitä, että dokumentin tosiaikaisen haun tehostamiseksi järjestely lisäksi käsittää patenttivaatimuksen 7 tai 8 mukaisen laitteiston (2-B).

15 10. Sähköisessä muodossa julkaistava dokumentti (3-20, 3-40), joka sisältää ainakin tekstiosuuden, tunnettu siitä, että dokumentin (3-20, 3-40) tosiaikaisen haun tehostamiseksi dokumentti (3-20, 3-40) käsittää täydennysosan (3-24), joka sisältää olennaisesti ainakin kyseisen dokumentin (3-20, 3-40) tekstiosuuden sisältämät sanat perusmuodoissaan alkuperäisessä järjestyksessä.

(57) Tiivistelmä

Sähköisesti julkaistavaan dokumenttiin (3-2) lisätään täydennysosana (3-24) ainakin dokumentin (3-2) tekstiosuuden sisältämät sanat perusmuodoissaan, alkuperäisessä järjestyksessä ja kunkin yhdyssanan kohdalla lisäksi yhdyssanan osien perusmuodot erillisinä sanoina. Täydennysosan lisääminen (2-20) saa aikaan sen, että hakupalvelu (1-A, 2-C) löytää seuraavan indeksoinnin (2-24) jälkeen keksinnön mukaisesti täydennetyn dokumentin (3-20, 3-40), vaikka alkuperäisessä dokumentissa sana ei esiintyisi lainkaan perusmuotoisena.

(Kuvio 3A)

FIG. 7

1-4

7-B

HAKUPALVELU

TE

DN-5

DOKUMENTIN
JULKAISIJA

1-2

WWW-OSOITE

7-4

IP-OSOITE

7-6

IP-OSOITE

7-8

WWW-SIVUT

7-10

ANNA-HAKUSANA

7-12

HAKUSANA(T)

1-14

ETS/DOKUMENTIT

7-16 LÖYDÖYT DOKUMENTIT

7-38

7-78

SELEÄ DOKUMENTIA

7-20

WWW-OSOITE

7-22

IP-OSOITE

7-24

IP-OSOITE

7-26

WWW-SIVUT

LOPPU

FIG. 2

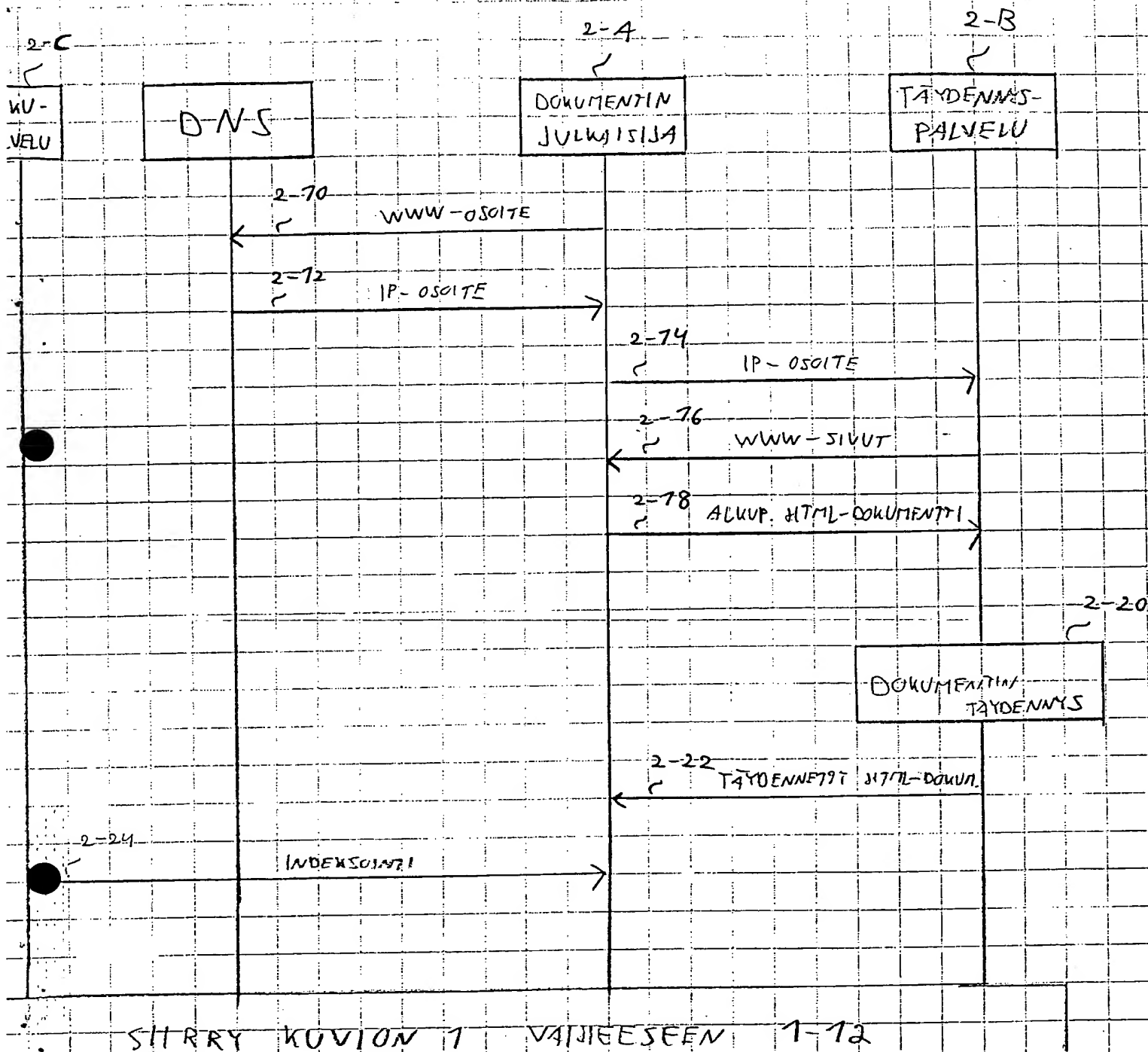


FIG. 3A

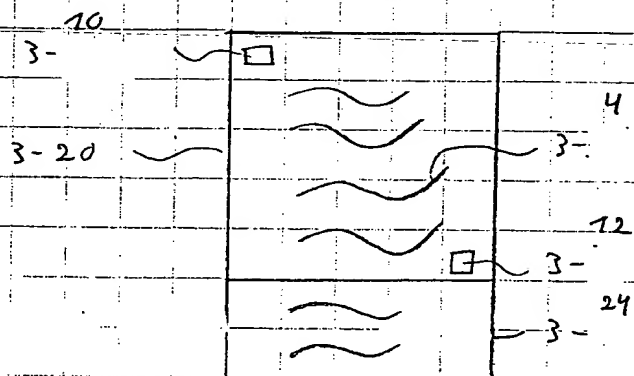
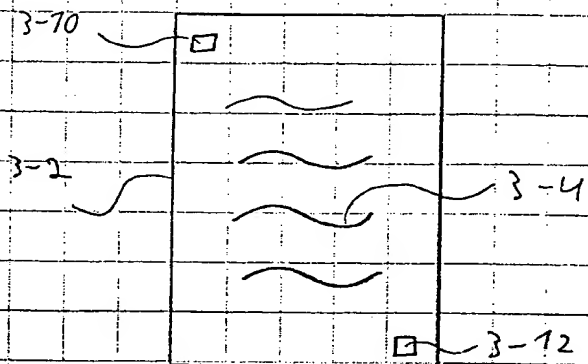


FIG. 3B

